27 The Correlative Brain: A Stream Segregation Model

Mounya Elhilali and Shihab Shamma

1 Introduction

The question of how everyday cluttered acoustic environments are parsed by the auditory system into separate streams is one of the most fundamental in perceptual science. Despite its importance, the study of its underlying neural mechanisms remains in its infancy; with a lack of general frameworks to account for *both* psychoacoustic and physiological experimental findings. Consequently, the few attempts at developing computational models of auditory stream segregation remain highly speculative. This in turn has considerably hindered the development of such capabilities in engineering systems such as automatic speech recognition, or sophisticated interfaces for communication aids (hearing aids, cochlear implants, speech-based human-computer interfaces).

In the current work, we present a mathematical model of auditory stream segregation, which accounts for both perceptual and neuronal findings of scene analysis. By closely coordinating with ongoing perceptual and physiological experiments, the proposed computational approach provides a rigorous framework for facilitating the integration of these results in a mathematical scheme of stream segregation, for developing effective algorithmic implementations to tackle the "cocktail party problem" in engineering applications, as well as generating new hypotheses to better understand the neural basis of active listening.

2 Framework and Foundation

2.1 Premise of the Model

Numerous studies have attempted to reveal the perceptual cues necessary and/or sufficient for sound segregation. Researchers have identified frequency separation, harmonicity, onset/offset synchrony, amplitude and

Institute for Systems Research & Department of Electrical and Computer Engineering, University of Maryland, College Park MD, USA, mounya@isr.umd.edu, sas@isr.umd.edu

Hearing – From Sensory Processing to Perception
B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey (Eds.)

frequency modulations, sound timbre and spatial location as the most prominent candidates for grouping cues in auditory streaming (Cooke and Ellis 2001). It is, however, becoming more evident that any sufficiently salient perceptual difference along *any auditory dimension* (at the periphery or central auditory stages) may lead to stream segregation.

On the biophysical level, our knowledge of neural properties particularly in the auditory cortex indicates that cortical responses (Spectro-Temporal Receptive Fields, STRFs) exhibit elaborate selectivity to spectral shapes, symmetry and dynamics of sound (Kowalski et al. 1996; Miller et al. 2002). This intricate mapping of acoustic waveforms into a *multidimensional* space suggests a role of the cortical circuitry in representing sounds in terms of auditory objects (Nelken 2004). Moreover, this organizational role is supported by the correspondence between time scales of cortical processing and the temporal dynamics of stream formation and auditory grouping.

In this study, we formalize these principles in a computational scheme that emphasizes two critical stages of stream segregation: (1) mapping sounds into a multi-dimensional feature space; (2) organizing sound features into temporally coherent streams. The *first* stage captures the mapping of acoustic patterns onto multiple auditory dimensions (tonotopic frequency, spectral timbre and bandwidth, harmonicity and common onsets). In this mapping, acoustic elements that evoke sufficiently non-overlapping activity patterns in the multi-dimensional representation space are deemed perceptually distinguishable and hence may potentially form distinct streams. We assume that these features are rapidly extracted and hence this mapping simulates "instantaneous" organization of sound elements (over short time windows; e.g. <200 ms), thus evoking the notion of *simultaneous* auditory grouping processes (Bregman 1990).

The *second* stage simulates the *sequential* nature of stream segregation. It highlights the principle that sound elements belonging to the same stream tend to *evolve together* in time. Conversely, temporally uncorrelated features are an indication of multiple streams or a disorganized acoustic scene. Identifying temporal coherence among multiple sequences of features requires integration of information over relatively long time periods (e.g. >300 ms), consistent with known dynamics of streaming-buildup. Therefore, the current model postulates that grouping features according to their levels of temporal coherence is a viable organizing principle underlying cortical mechanisms in sound segregation.

2.2 Stage 1: Multi-dimensional Cortical Representation

Current understanding of auditory cortical processing inspires our model for the multi-dimensional representation of sound. The model takes in as input an auditory spectrogram, and effectively performs a wavelet decomposition using a bank of linear spectro-temporal receptive fields (STRFs). The

analysis proceeds in two steps (as detailed in Chi et al. 2005): (i) a *spectral* step that maps each incoming spectral slice into a 2D frequency-scale representation. It is implemented by convolving the time-frequency spectrogram $y(t,x)$ with a complex-valued spectral receptive field SRF, parametrized by spectral tuning Ω_c and characteristic phase ϕ_c; (ii) a *temporal* step in which the time-sequence from each frequency-scale combination (channel) is convolved with a temporal receptive field TRF to produce the final 4D cortical mapping *r*. Each temporal filter is characterized by its modulation rate ω_c and phase θ_c. This cortical mapping is depicted in Fig. 1A, and can be captured by

$$\begin{aligned} s\,(t, x; \Omega_c, \phi_c) &= y\,(t, x) *_x \mathcal{SRF}(x; \Omega_c, \phi_c) \\ r(t, x; \omega_c, \Omega_c, \theta_c, \phi_c) &= s\,(t, x; \Omega_c, \phi_c) *_t \mathcal{TRF}(t; \omega_c, \theta_c) \end{aligned} \tag{1}$$

We choose the model's parameters to be consistent with cortical response properties, spanning the range Γ=[0.5–4] peaks/octave spectrally and Ψ = [1–30] Hz temporally. Clearly, other feature dimensions (such as spatial location and pitch) can supplement this multidimensional representation as needed.

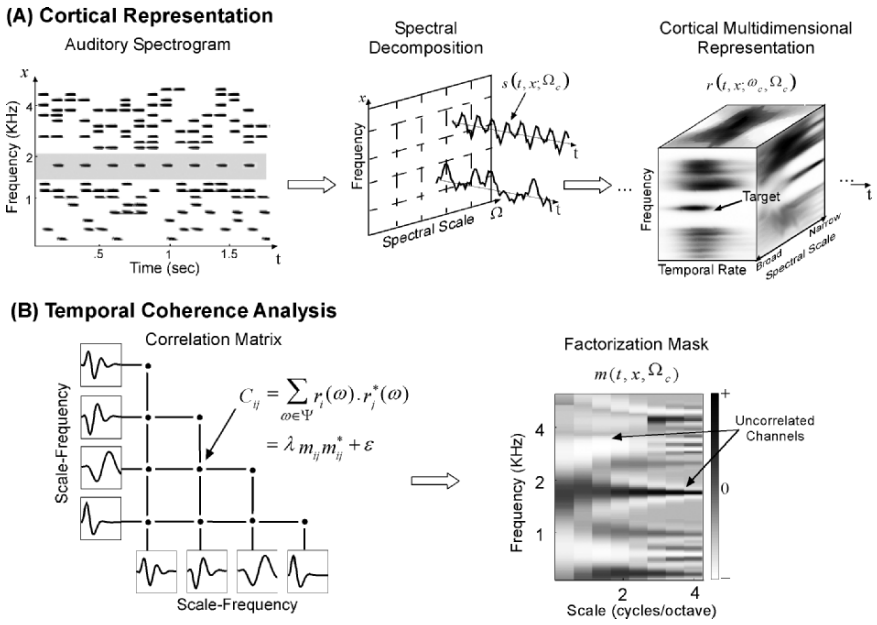

Fig. 1 A,B Schematic of stream segregation model

2.3 Stage 2: Temporal Coherence Analysis

The essential function of this stage is twofold: (i) estimate a pair-wise correlation matrix (C) among all scale-frequency channels, and then (ii) determine from it the optimal factorization of the spectrogram into two streams (foreground and background) such that responses within each stream are maximally coherent.

The correlation is derived from an instantaneous coincidence match between all pairs of frequency-scale channels integrated over time. Given that TRF filters provide an analysis over multiple time windows, this step is equivalent to an instantaneous pair-wise correlation across channels summed over rate filters (Fig. 1B):

$$\textit{Correlation Matrix} = \int s_i(t)\, s_j(t)\, dt \simeq \sum_{\omega \in \psi} r_i(\omega)\, r_j^*(\omega) \triangleq C_{ij} \qquad (2)$$

where (*) denotes the complex-conjugate. We can find the "optimal" factorization of this matrix into two uncorrelated streams, by determining the direction of maximal incoherence between the incoming stimulus patterns. Such a factorization is accomplished by a principal component analysis of the correlation matrix C (Golub and Van Loan 1996), where the principal eigenvector corresponds to a map labeling channels as positively or negatively correlated entries. The value of its corresponding eigenvalue reflects the degree to which the matrix C is decomposable into two uncorrelated sets, and hence reflects how 'streamable' the input is.

2.4 Computing the Two Streams

Therefore, the computational algorithm for factorizing the matrix C is as follows:

1. At each time step, the matrix $C(t)$ is computed from the cortical representation as in Eq. (2). The correlation matrix keeps evolving as the cortical output $r(t)$ changes over time. However for stationary stimuli, the correlation pattern reaches a stable point after a buildup period.
2. Given its hermitian nature (since it is a correlation matrix), C can be expressed as $C = \lambda m m^{\dagger} + \varepsilon$, where m is the principal eigenvector of C, λ its corresponding eigenvalue, and $\varepsilon(t)$ the residual energy in C not accounted for by the outer-product of m. ($\dagger$) denotes the hermitian transpose. The ratio of λ^2 to the total energy in C corresponds to the proportion of the correlation matrix accounted for by its best factorization m. This ratio is an indicator of the separablity of the matrix C, and hence the streamability of the sound.

The principal eigenvector m can be viewed as a 'mask', which can differentially shape the scale-frequency input pattern at any given time instant. This mask

consists of a map of weights that positively scales channels with a common orientation and suppresses channels in the opposite direction. Effectively, m (and its complement $1–m$) acts as a "filter" through which we can produce the foreground (and background) stream.

3 Simulation Results

The model was tested on several classic stream segregation conditions to demonstrate its ability to emulate known percepts as reported by human subjects. The first row in Fig. 2 illustrates results of the classic alternating tone paradigm (Bregman 1990). The leftmost panel shows the mask profile m for this stimulus. Given its stationary nature, the matrix C stabilizes rapidly, and its factorization m reveals that the energy in channel A (low tone) is temporally anti-correlated with channel B (high tone), and hence should belong to a different stream.

The second row of Fig. 2 depicts simulation results for a target tone in a multi-tone background, commonly used in Informational Masking (IM) tests. This stimulus is the focus of the remainder of this study, where we attempt to use the model to account for perceptual and physiological results using the same paradigm.

The right lower panels of Fig. 2 show the outcome of applying the mask m to the IM spectrogram. As the correlation pattern builds up in time, the target tone is flagged as temporally un-correlated with the background tones, and hence is slowly suppressed in the left stream. Given the random nature of the background, some maskers are occasionally labeled as weakly correlated with the target. This explains why the target stream has a weak contribution from the maskers.

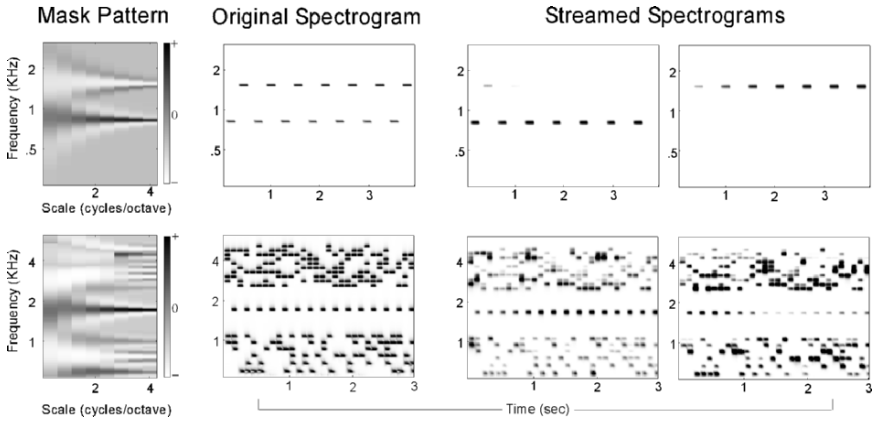

Fig. 2 Model simulations

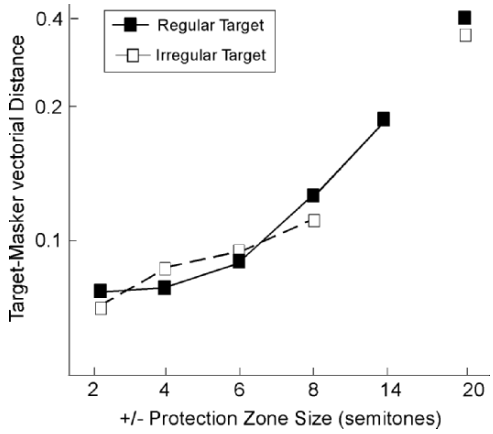

Fig. 3 Predicted target detection

4 Perceptual Measures

To validate the simulation results against human perception with IM stimuli, we derived a measure of how detectable the target is, based on our mask profile. The measure quantifies the mean vectorial distance between the complex-valued energy of *m* at the target channel, and energy in any other masker channel. Figure 3 illustrates the change in this distance *d* as the protection zone separating the target from the maskers varies. In accord with findings from psychoacoustic tests (Micheyl et al. 2007), the trend in this distance plot reveals that unmasking effects of the target depend on the size of the spectral protective region around the target tone. Additionally, the model reveals that temporal regularity of the target does not seem to be a critical cue for target detection. The open symbols in Fig. 3 demonstrate that regular targets or roved irregular targets (average of one target every two masker bursts) yield virtually similar distance values; and hence result in similar unmasking levels, as shown by perceptual findings in (Micheyl et al. 2007).

5 Physiological Correlates

In addition to mimicking human perceptual performance, the model enables us to explore neural correlates of streaming and attention, as observed in physiological studies using the same IM paradigm (Yin et al. 2007). To do so, we add more biological realism to the model by incorporating a stage of neuronal adaptation, simulated via mechanisms of synaptic depression known to

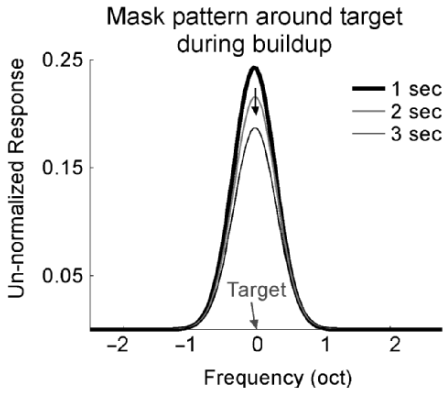

Fig. 4 Gain change of target tuning curve as predicted by model's *mask* response

operate at the thalamo-cortical projections (as described in Elhilali et al. 2003). This stage shapes the energy pattern of each channel at the input of the cortical model by effectively adapting its activity in a nonlinear fashion. This neuronal adaptation has been explored as a potential mechanism underlying observed tuning curve changes in *naïve* or *non-behaving* animals presented with streaming-like paradigms (Yin et al. 2007). Consistent with these speculation, our simulations reveal a drop in tuning curve gain during the buildup period (Fig. 4). These tuning curves are obtained by weighting the model's spectral receptive fields (SRF*)* region around the target tone with its corresponding mask profile *m* at different time epochs of the stimulus.

By contrast, simulating behavioral shifts in *trained* animals has to evoke top-down attentional mechanisms which would for instance modulate the weights of the cortical map, by emphasizing the STRF regions associated with the task at hand. Specifically, when a trained animal is performing a detection task of a single tone surrounded by broadly distributed masker tones (referred to as Task 2 in Yin et al. 2007), a potential mechanism at play is learning to promote narrowly tuned neuronal ensembles so as to focus on a single target tone. Such consistent attentional emphasis can be simulated by applying a high-pass to the scale dimension in Fig. 1, hence amplifying the response from the high scales (i.e., narrowband) region. Conversely, when the animal learns to attend to the broadband masker background tones (Task 1), it could potentially emphasize activity in the broadband region. We simulate this situation by a low-pass along the scale dimension. The effect of these task dependencies is illustrated in Fig. 5, which depicts the changing bandwidth of a tuning curve *during* the performance of these two tasks, as shown in physiological findings in (Yin et al. 2007).

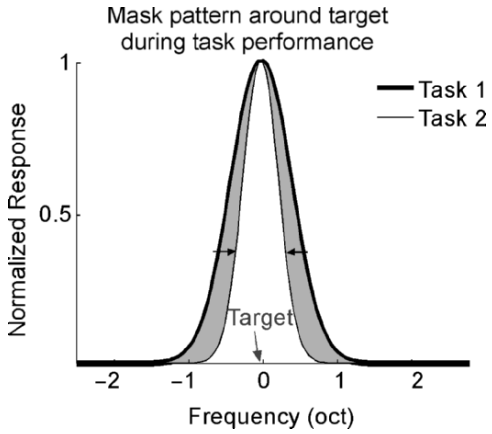

Fig. 5 Bandwidth changes of target tuning curve as predicted by model's *mask* response during task performance

6 Final Remarks

We have demonstrated that analysis of response *coherence* in a model of auditory cortical processing can account for the perceptual organization of sound streams. While response *coherence* emerges as the key overarching organizational principle, its computational implementation can take different but essentially equivalent forms. For instance, this paper focused on the correlation matrix C and its factorization as the vehicle for the analysis. Alternatively, a focus on predicting response consistency within different streams results in a Kalman filtering interpretation (Elhilali and Shamma 2006). Ongoing and future investigations must also incorporate biologically plausible adaptive mechanisms to account for the observed effects of behavior on cortical responses during streaming.

Acknowledgment. This work is supported by CSRNS RO1 AG02757301, AFOSR and SWRI.

References

Bregman A (1990) Auditory scene analysis. MIT Press

Chi T, Ru P, Shamma S (2005) Multiresolution spectrotemporal analysis of complex sounds, J Acoust Soc Am 118:887–906

Cooke M, Ellis D (2001) The auditory organization of speech and other sources in listeners and computational models. Speech Commun 35:141–177

Elhilali M, Shamma S (2006) A biologically-inspired approach to the cocktail party problem. Proc ICASSP

Elhilali M, Fritz J, Klein D, Simon J, Shamma S (2003) Dynamics of precise spiking in primary auditory cortex. J Neurosci 24(5):1159–1172

Golub G, Van Loan C (1996) Matrix computations, 3rd edn. Johns Hopkins Univ Press
Kowalski N, Depireux D, Shamma S (1996) Analysis of dynamic spectra in ferret primary auditory cortex. J Neurophysiol 76:3503–3523
Micheyl C, Oxenham A, Shamma S (2007) Detection of repeating targets in random multi-tone backgrounds: perceptual mechanisms. Current volume
Miller L, Escabi M, Read H, Schreiner C (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol 87(1):516–527
Nelken I (2004) Processing of complex stimuli and natural scenes in the auditory cortex. Curr Opin Neurobiol 14:474–480
Yin P, Ma L, Elhilali M, Fritz J, Shamma S (2007) Neural correlates of attention during streaming. Current volume

Comment by Yost

The discussion following your excellent talk, underscored what I think can be an important distinction. I do not believe that 'streaming' and 'source' segregation are always the same thing. That is, in your A-B example two sounds (A and B) that occur at the same time can be perceived as coming from two sources, but they may not be perceived as being a continuation of the sources perceived at a different point in time – segregation may occur but streaming did not. From my perspective, streaming is a form of source segregation that involves an element of continuity over time. Or, put another way streaming is an example of source segregation, but they are not the same thing. You presented your model as a stream segregation model, but it appears that with the proper time constants it might also be used for source segregation in the absence of perceived continuity from stimulus presentation to stimulus presentation. For instance, with a very short time constant the model might be able to account for the segregation of two different transients that occurred at the same time. Is this correct?

Reply

I agree with your argument about the use of the model (namely the first stage of a sound multi-feature representation) as a scheme for segregating sound components present in the environment at any instant in time. In this multi-dimensional representation, acoustic elements that evoke sufficiently non-overlapping activity patterns in the feature space are deemed perceptually distinguishable and can hence be perceived as individual components in a complex scene at any instant in time. This sound representation builds up over tens of milliseconds (<150 ms), but reflects the instantaneous segregation of an acoustic scene. In contrast, the temporal coherence stage of the model reflects the dynamic nature of stream segregation as it builds up over time requiring information integration over few hundred milliseconds. Hence, as expressed in your comment, the 'instantaneous' elements parsed in the first stage might or might not evolve to a percept of segregated streams,

depending on whether they maintain a coherent evolution over time from one stimulus presentation to the next stimulus presentation.

My only disagreement with your statement is the use of the term 'source' segregation, because that expression reflects more the physical cause of a sound, and not necessarily our perception of the individual components of a sound. Hence, I would prefer to call this instantaneous segregation a parsing of the scene into its constituent elements; which allow us at every instant of time to perceive different elements present in the environment (which you called sources).

Comment by Divenyi

I don't think anybody would argue that the basic premise of streaming is source perception. When a sequence is perceived as two streams, it is that we attribute the two alternating sounds as coming from two different sources. Conversely, when the sequence is perceived as a single stream, we attribute the whole sequence to a single source. So, when two sounds that are segregated into two streams are now played simultaneously and repeated over a longer period, they are grouped together by virtue of their shared temporal properties. I think that the listener will end up considering the ensemble as being produced by a single source, just like a consonant burst is considered as coming from one source regardless of how many disparate spectral patches it may consist of. In music, too, a repeated chord, no matter how complex, will be considered as the same repeated event. Would not it be preferable that the correlation metric you propose would indicate the number of sources instead?

Reply

I do agree that the percepts that arise from many acoustic scenes do not necessarily reflect the actual physical sound sources present in the environment. However, I would disagree with your statement that the premise of streaming is source separation. Rather, I would agree with Bregman's definition of stream where he argued to reserve the word 'stream' for the perceptual representation, and the word 'sound' or 'source' for the physical cause. Aside from the nomenclature issue, I completely agree with your argument.

As far as the use of the model for indicating the number of streams (or 'sources') in the scene, we can definitely expand our formulation to incorporate information from the second and higher principal dimensions of the coherence correlation matrix C (after performing the matrix factorization). These additional degrees of freedom can indicate the presence of a third or fourth stream whose components are highly correlated amongst themselves. We have not yet explored this extension of the model in the current study, but will try to incorporate it in alternative implementations of the model.